Necesarry libraries

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.1.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

Loading dataset

```
dataset <- read_csv("preprocessed_data.csv")
```

```
## Rows: 1460 Columns: 81
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (43): MSZoning, Street, Alley, LotShape, LandContour, Utilities, LotConf...
## dbl (38): Id, MSSubClass, LotFrontage, LotArea, OverallQual, OverallCond, Ye...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Pitanje 1. "Određuje li oblik zemljišne čestice broj katova kuće?"
```

Postavljamo hipoteze:

Ho: broj katova kuće ne ovisi o obliku zemljišne čestice H1: broj katova kuće ovisi o obliku zemljišne čestice

Učitavamo dataset u varijablu df sa kojom ćemo dalje raditi.

```
df <- dataset
```

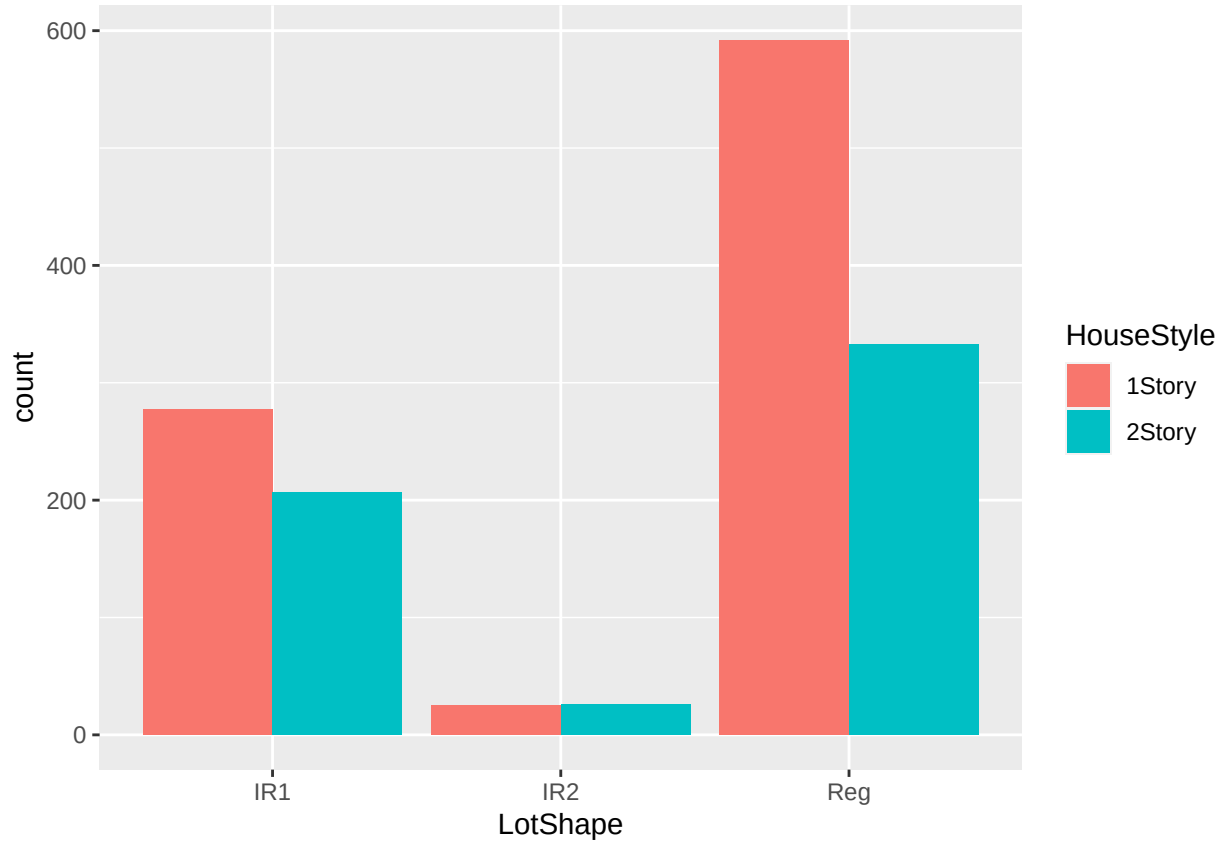
Grupiramo vrijednosti podataka u stupcu. Stambene objekte koji imaju 1.5 kat tretiramo kao da imaju 1 kat, a objekte koji imaju 2.5 kata tretiramo kao da imaju 2 kata.

```
df$HouseStyle <- gsub("1.5Fin", "1Story", df$HouseStyle)
df$HouseStyle <- gsub("1.5Unf", "1Story", df$HouseStyle)
df$HouseStyle <- gsub("2.5Fin", "2Story", df$HouseStyle)
df$HouseStyle <- gsub("2.5Unf", "2Story", df$HouseStyle)
df$HouseStyle <- gsub("SFoyer", "2Story", df$HouseStyle)
df$HouseStyle <- gsub("SLvl", "2Story", df$HouseStyle)
df$LotShape <- gsub("IR3", "IR2", df$LotShape)
```

```
df
```

```
## # A tibble: 1,460 x 81
##       Id MSSubClass MSZon~1 LotFr~2 LotArea Street Alley LotSh~3 LandC~4 Utili~5
##    <dbl>      <dbl> <chr>     <dbl>   <dbl> <chr>  <chr> <chr>   <chr>   <chr>
##  1     1         60 RL           65    8450 Pave   <NA>  Reg     Lvl     AllPub
##  2     2         20 RL           80    9600 Pave   <NA>  Reg     Lvl     AllPub
##  3     3         60 RL           68   11250 Pave   <NA>  IR1     Lvl     AllPub
##  4     4         70 RL           60    9550 Pave   <NA>  IR1     Lvl     AllPub
##  5     5         60 RL           84   14260 Pave   <NA>  IR1     Lvl     AllPub
##  6     6         50 RL           85   14115 Pave   <NA>  IR1     Lvl     AllPub
##  7     7         20 RL           75   10084 Pave   <NA>  Reg     Lvl     AllPub
##  8     8         60 RL           NA   10382 Pave   <NA>  IR1     Lvl     AllPub
##  9     9         50 RM           51    6120 Pave   <NA>  Reg     Lvl     AllPub
## 10    10        190 RL           50    7420 Pave   <NA>  Reg     Lvl     AllPub
## # ... with 1,450 more rows, 71 more variables: LotConfig <chr>,
## #   LandSlope <chr>, Neighborhood <chr>, Condition1 <chr>, Condition2 <chr>,
## #   BldgType <chr>, HouseStyle <chr>, OverallQual <dbl>, OverallCond <dbl>,
## #   YearBuilt <dbl>, YearRemodAdd <dbl>, RoofStyle <chr>, RoofMatl <chr>,
## #   Exterior1st <chr>, Exterior2nd <chr>, MasVnrType <chr>, MasVnrArea <dbl>,
## #   ExterQual <chr>, ExterCond <chr>, Foundation <chr>, BsmtQual <chr>,
## #   BsmtCond <chr>, BsmtExposure <chr>, BsmtFinType1 <chr>, ...
```

```
ggplot(df, aes(x = LotShape, fill = HouseStyle)) +
  geom_bar(position = "dodge")
```



Napravit ćemo kontingencijsku tablicu i dodati joj marginalne vrijednosti.

```
table <- table(df$LotShape, df$HouseStyle)

margins_tbl = addmargins(table)
print(margins_tbl)
```

```
##
##       1Story 2Story  Sum
##   IR1    277    207  484
##   IR2     25     26   51
##   Reg    592    333  925
##   Sum    894    566 1460
```

`chisq.test()` se može izvršiti samo ako očekivana frekvencija pojedinog razreda iznosi najmanje 5. Pretpostavka testa je da je ovo uvjet zadovoljen, stoga se prije provođenja testa mora provjeriti da li je očekivana frekvencija pojedinog razreda veća ili jednaka 5.

```
for (col_names in colnames(margins_tbl)){
  for (row_names in rownames(margins_tbl)){
    if (!(row_names == 'Sum' | col_names == 'Sum') ){
      cat('Očekivane frekvencije za ',col_names,'-',row_names,': ',(margins_tbl[row_names,'Sum'] * margi
    }
  }
}
```

```
## Očekivane frekvencije za  1Story - IR1 :  296.3671
## Očekivane frekvencije za  1Story - IR2 :  31.22877
## Očekivane frekvencije za  1Story - Reg :  566.4041
## Očekivane frekvencije za  2Story - IR1 :  187.6329
## Očekivane frekvencije za  2Story - IR2 :  19.77123
## Očekivane frekvencije za  2Story - Reg :  358.5959
```

Sve očekivane frekvencije su veće od 5. Možemo nastaviti sa`chisq.test()` testom. Prethodno spomenute hipoteze testiramo testom nezavisnoti pomoću chi testa.

```
test = chisq.test(df$LotShape, df$HouseStyle, simulate.p.value = TRUE, B = 1000)
```

$p < 0.05$ Na temelju p vrijednosti odbacujemo hipotezu "H0:broj katova kuće ne ovisi o obliku zemljišne čestice". Prihvaćamo alternativnu hipotezu "H1: broj katova kuće ovisi o obliku zemljišne čestice" i zaključujemo da broj katova kuće ovisi o obliku zemljišne čestice.